整数計画法による文字列の集合上の確率分布に対する中央文字列と中心文字列

AUTHOR(S):

林田, 守広

CITATION:

林田, 守広. 整数計画法による文字列の集合上の確率分布に対する中央文字列と中心文字列. 京都大学化学研究所スーパーコンピュータシステム研究成果報告書 2016, 2015: 14-14

ISSUE DATE:

2016-03

URL:

http://hdl.handle.net/2433/214401

平成27年度　　京都大学化学研究所　スーパーコンピュータシステム　利用報告書

整数計画法による文字列の集合上の確率分布に対する中央文字列と中心文字列
Integer linear programming approach to median and center strings for a probability distribution on a set of strings

京都大学化学研究所バイオインフォマティクスセンター数理生物情報　林田守広

研究成果概要

文字列の集合上の確率分布に対するレーベンシュタイン距離での中央文字列あるいは中心文字列を求める問題は NP 困難であり難しい. 本研究では整数計画問題による定式化を提案した. 有限個の文字からなるアルファベットを A とし, 文字列の集合 A*上の確率分布を p(s)とする. 二つの文字列 s, t の間のレーベンシュタイン距離を d(s, t)とするとき, 中央文字列 m はすべての文字列 s に対して p(s)d(m, s)の和を最小とする A*の元である. また中心文字列 c はすべての文字列 s における p(s)d(c, s)の最大値を最小化する文字列であると定義される.

レーベンシュタイン距離は, 動的計画法を用いて与えられた文字列の長さの多項式時間で計算可能であるが, 整数計画問題における線形式として直接記述することは困難である. そこで, 各文字の間での挿入, 削除, 置換の有無を整数計画問題中の変数として表現した. 中央文字列を求める問題では, これらの変数を使って p(s)d(m, s)の和を表現し, これを最小化する問題として定式化した. 中心文字列を求める問題では, p(s)d(c, s)がある値 h 以下であるという制約を条件として加え, h を最小化する問題として定式化した. これらの整数計画問題は厳密に中央文字列および中心文字列を解とする一方, 文字列の探索空間は膨大である. そのため高い確率をもつ文字列どうしのレーベンシュタイン距離が小さい場合に, 厳密解法に制約条件を加えることで探索空間を縮小し近似解を求める定式化の方法も提案した.

DNA あるいは RNA の塩基配列は 4 種類の文字からなっていることより|A|=4 としていくつかの A*上の確率分布を用いて計算時間を複数回計測しその平均と分散を計算した. 中央文字列, 中心文字列を求める問題は NP 困難であるため, 正の確率をもつ文字列の数また文字列長の増加について, 提案する厳密解法では急激に計算時間が増加した. 一方, 近似解法では緩やかな増加に留まった.

発表論文（謝辞あり）

発表論文（謝辞なし）

Hayashida, M, and Koyano, H, Integer linear programming approach to median and center strings for a probability distribution on a set of strings, *The seventh International Conference on Bioinformatics Models, Methods and Algorithms*, Feb. 2016.